Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators

Shereen Oraby[1], **Lena Reed**[1], **Shubhangi Tandon**[1],
Sharath T.S.[1], **Stephanie Lukin**[2], **and Marilyn Walker**[1]
[1]Natural Language and Dialogue Systems Lab, University of California, Santa Cruz
[2]U.S. Army Research Laboratory, Los Angeles, CA
{soraby,lireed,shtandon,sturuvek,mawalker}@ucsc.edu
stephanie.m.lukin.civ@mail.mil

Abstract

Natural language generators for task-oriented dialogue must effectively realize system dialogue actions and their associated semantics. In many applications, it is also desirable for generators to control the style of an utterance. To date, work on task-oriented neural generation has primarily focused on semantic fidelity rather than achieving stylistic goals, while work on style has been done in contexts where it is difficult to measure content preservation. Here we present three different sequence-to-sequence models and carefully test how well they disentangle content and style. We use a statistical generator, PERSONAGE, to synthesize a new corpus of over 88,000 restaurant domain utterances whose style varies according to models of personality, giving us total control over both the semantic content and the stylistic variation in the training data. We then vary the amount of explicit stylistic supervision given to the three models. We show that our most explicit model can simultaneously achieve high fidelity to both semantic and stylistic goals: this model adds a context vector of 36 stylistic parameters as input to the hidden state of the encoder at each time step, showing the benefits of explicit stylistic supervision, even when the amount of training data is large.

1 Introduction

The primary aim of natural language generators (NLGs) for task-oriented dialogue is to effectively realize system dialogue actions and their associated content parameters. This requires training data that allows the NLG to learn how to map semantic representations for system dialogue acts to one or more possible outputs (see Figure 1, (Novikova et al., 2016)). Because neural generators often make semantic errors such as deleting, repeating or hallucinating content, to date previous work on task-oriented neural generation has primarily focused on faithfully rendering the meaning of the system's dialogue act (Dusek and Jurcícek, 2016b; Lampouras and Vlachos, 2016; Mei et al., 2015; Wen et al., 2015).

INFORM(NAME[THE EAGLE], EATTYPE[COFFEE SHOP], FOOD[ENGLISH], PRICERANGE[HIGH], CUSTOMERRATING[AVERAGE], AREA[CITY CENTRE], FAMILYFRIENDLY[YES], NEAR[BURGER KING])

The three star coffee shop, The Eagle, located near Burger King, gives families a high priced option for English food in the city centre.

Let's see what we can find on The Eagle. Right, The Eagle is a coffee shop with a somewhat average rating. The Eagle is kid friendly, also it's an English restaurant and expensive, also it is near Burger King in the city centre, you see?

Figure 1: Dialogue Act Meaning Representation (MR) with content parameters and outputs

However, in many applications it is also desirable for generators to control the style of an utterance independently of its content. For example, in Figure 1, the first output uses more formal language and complex syntactic structures, as one might see in written language, while the other uses simpler syntax and pragmatic markers characteristic of oral language (Biber, 1991). In this paper, we create several different sequence-to-sequence models and compare how well they can disentangle content and style. Controlling the style of the output requires disentangling the content from the style, but previous neural models aimed at achieving stylistic goals have not focused on task-oriented dialogue where specific semantic attributes and values must be communicated (as per the MR in Figure 1), and where semantic fi-

Proceedings of the SIGDIAL 2018 Conference, pages 180–190,
Melbourne, Australia, 12-14 July 2018. ©2018 Association for Computational Linguistics

Personalities	Realization
MEANING REPRESENTATION (MR)	NAME[FITZBILLIES], EATTYPE[PUB], FOOD[ITALIAN], PRICERANGE[MODERATE], CUSTOMER RATING[DECENT], AREA[RIVERSIDE], FAMILYFRIENDLY[NO],NEAR['THE SORRENTO']
NO-AGG/ NO-PRAG	Fitzbillies is a pub. Fitzbillies has a decent rating. Fitzbillies is moderately priced. Fitzbillies is in riverside. Fitzbillies is an Italian restaurant. Fitzbillies is not family friendly. Fitzbillies is near The Sorrento.
AGREEABLE	Let's see what we can find on Fitzbillies. I see, well it is a pub with a decent rating, also it is an Italian restaurant in riverside and moderately priced near The Sorrento, also it isn't family friendly, you see?
DISAGREEABLE	I mean, everybody knows that moderately priced Fitzbillies is in riverside with a decent rating. It's near The Sorrento. It isn't family friendly. It is an Italian place. It is a pub.
CONSCIENTIOUS	Let's see what we can find on Fitzbillies. I see, well it is a pub with a decent rating, it isn't kid friendly and it's moderately priced near The Sorrento and an Italian restaurant in riverside.
UNCONSCIENTIOUS	Oh god yeah, I don't know. Fitzbillies is a pub with a decent rating, also it is moderately priced near The Sorrento and an Italian place in riverside and it isn't kid friendly.
EXTRAVERT	Basically, Fitzbillies is an Italian place near The Sorrento and actually moderately priced in riverside, it has a decent rating, it isn't kid friendly and it's a pub, you know.

Table 1: Sample neural model output realizations for the same MR for PERSONAGE's personalities

delity can be precisely measured.[1]

One of the main challenges is the lack of parallel corpora realizing the same content with different styles. Thus we create a large, novel parallel corpus with specific style parameters and specific semantics, by using an existing statistical generator, PERSONAGE (Mairesse and Walker, 2010), to synthesize over 88,000 utterances in the restaurant domain that vary in style according to psycholinguistic models of personality.[2] PERSONAGE can generate a very large number of stylistic variations for any given dialogue act, thus yielding, to our knowledge, the largest style-varied NLG training corpus in existence. The strength of this new corpus is that: (1) we can use the PERSONAGE generator to generate as much training data as we want; (2) it allows us to systematically vary a specific set of stylistic parameters and the network architectures; (3) it allows us to systematically test the ability of different models to generate outputs that faithfully realize both the style and content of the training data.[3]

We develop novel neural models that vary the amount of explicit stylistic supervision given to the network, and we explore, for the first time, explicit control of multiple interacting stylistic parameters. We show that the no-supervision (NO-SUP) model, a baseline sequence-to-sequence model (Sutskever et al., 2014; Dusek and Jurcícek, 2016b), produces semantically correct outputs, but eliminates much of the stylistic variation that it saw in the training data. MODEL_TOKEN provides minimal supervision by allocating a latent variable in the encoding as a label for each style, similar to the use of language labels in machine translation (Johnson et al., 2017). This model learns to generate coherent and stylistically varied output without explicit exposure to language rules, but makes more semantic errors. MODEL_CONTEXT adds another layer to provide an additional encoding of individual stylistic parameters to the network. We show that it performs best on both measures of semantic fidelity and stylistic variation. The results suggest that neural architectures can benefit from explicit stylistic supervision, even with a large training set.

2 Corpus Creation

We aim to systematically create a corpus that can be used to test how different neural architectures affect the ability of the trained model to disentangle style from content, and faithfully produce semantically correct utterances that vary style. We use PERSONAGE, an existing statistical generator: due to space, we briefly explain how it works, referring the interested reader to Mairesse and Walker (2010, 2011) for details.

PERSONAGE requires as input: (1) a meaning representation (MR) of a dialogue act and its content parameters, and (2) a parameter file that tells it how frequently to use each of its stylistic parameters. Sample model outputs are shown in the second row of Figure 1 and in Table 1, illustrating some stylistic variations PERSONAGE produces.

To generate our novel corpus, we utilize the

[1] We leave a detailed review of related work to Section 6.

[2] Our stylistic variation for NLG corpus is available at: nlds.soe.ucsc.edu/stylistic-variation-nlg

[3] Section 4 quantifies the naturalness of PERSONAGE outputs.

MRs from the E2E Generation Challenge.[4] The MR in Figure 1 illustrates **all 8** available attributes. We added a dictionary entry for each attribute to PERSONAGE so that it can express that attribute.[5] These dictionary entries are syntactic representations for very simple sentences: the NO-AGG/NO-PRAG row of Table 1 shows a sample realization of each attribute in its own sentence based on its dictionary entry.

	Number of Attributes in MR					
Dataset	**3**	**4**	**5**	**6**	**7**	**8**
TRAIN	0.13	0.30	0.29	0.22	0.06	0.01
TEST	0.02	0.04	0.06	0.15	0.35	0.37

Table 2: Percentage of the MRs in training and test in terms of number of attributes in the MR

We took advantage of the setup of the E2E Generation Challenge and used their MRs, exactly duplicating their split between training, dev and test MRs, because they ensured that the dev and test MRs had not been seen in training. The frequencies of longer utterances (more attribute MRs) vary across train and test, with actual distributions in Table 2, showing how the test set was designed to be challenging, while the test set in Wen et al. (2015) averages less than 2 attributes per MR (Nayak et al., 2017). We combine their dev and training MRs resulting in 3784 unique MRs in the training set, and generate 17,771 reference utterances per personality for a training set size of 88,855 utterances. The test set consists of 278 unique MRs and we generate 5 references per personality for a test size of 1,390 utterances.

The experiments are based on two types of parameters provided with PERSONAGE: aggregation parameters and pragmatic parameters.[6] The NO-AGG/NO-PRAG row of Table 1 shows what PERSONAGE would output if it did not use any of its stylistic parameters. The top half of Table 3 illustrates the aggregation parameters: these parameters control how the NLG combines attributes into sentences, e.g., whether it tries to create complex sentences by combining attributes into phrases and what types of combination operations it uses. The pragmatic operators are shown in the second half of Table 3. Each parameter value can be set to `high`, `low`, or `don't care`.

Attribute	**Example**
AGGREGATION OPERATIONS	
PERIOD	*X serves Y. It is in Z.*
"WITH" CUE	*X is in Y, with Z.*
CONJUNCTION	*X is Y and it is Z. & X is Y, it is Z.*
ALL MERGE	*X is Y, W and Z & X is Y in Z*
"ALSO" CUE	*X has Y, also it has Z.*
PRAGMATIC MARKERS	
ACK_DEFINITIVE	*right, ok*
ACK_JUSTIFICATION	*I see, well*
ACK_YEAH	*yeah*
CONFIRMATION	*let's see what we can find on X,* *let's see ....., did you say X?*
INITIAL REJECTION	*mmm, I'm not sure, I don't know.*
COMPETENCE MIT.	*come on, obviously, everybody knows that*
FILLED PAUSE STATIVE	*err, I mean, mmhm*
DOWN_KIND_OF	*kind of*
DOWN_LIKE	*like*
DOWN_AROUND	*around*
EXCLAIM	*!*
INDICATE SURPRISE	*oh*
GENERAL SOFTENER	*sort of, somewhat, quite, rather*
DOWN_SUBORD	*I think that, I guess*
EMPHASIZER	*really, basically, actually, just*
EMPH_YOU_KNOW	*you know*
EXPLETIVES	*oh god, damn, oh gosh, darn*
IN GROUP MARKER	*pal, mate, buddy, friend*
TAG QUESTION	*alright?, you see? ok?*

Table 3: Aggregation and Pragmatic Operations

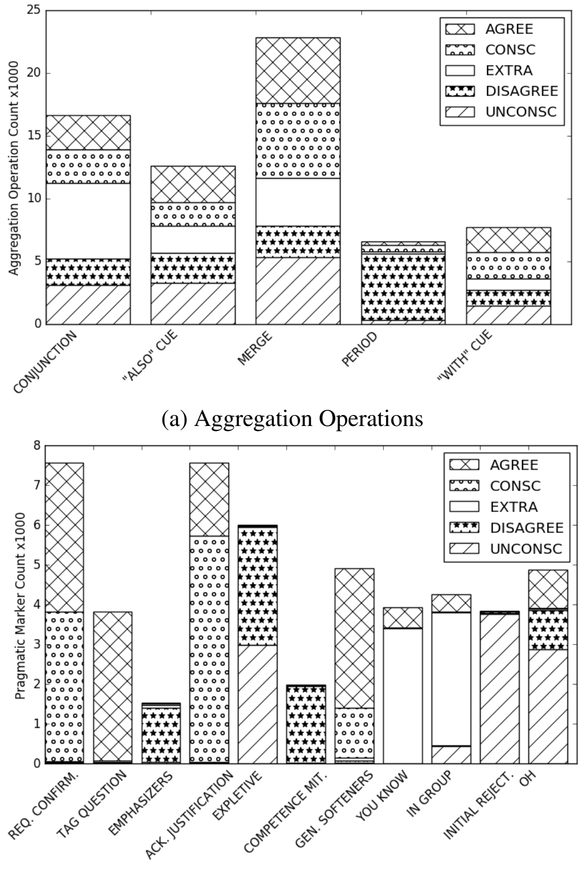
To use PERSONAGE to create training data mapping the same MR to multiple personality-based variants, we set **values** for **all** of the parameters in Table 3 using the stylistic models defined by Mairesse and Walker (2010) for the following Big Five personality traits: agreeable, disagreeable, conscientiousness, unconscientiousness, and extravert. Figure 2 shows that each personality produces data that represents a stylistically distinct distribution. These models are probabilistic and specified values are automatically broadened within a range, thus each model can produce 10's of variations for each MR. Note that while each personality type distribution can be characterized by a single stylistic label (the personality), Figure 2 illustrates that each distribution is characterized by multiple interacting stylistic parameters.

Each parameter modifies the linguistic structure in order to create distributionally different subcorpora. To see the effect of each personality using a different set of aggregation operators, cross-reference the aggregation operations in Table 3 with an examination of the outputs in Table 1. The

[4] `http://www.macs.hw.ac.uk/InteractionLab/E2E/`

[5] PERSONAGE supports a one-to-many mapping from attributes to elementary syntactic structures for expressing that attribute, but here we use only one dictionary entry. PERSONAGE also allows for discourse relations such as justification or contrast to hold between content items, but the E2E MRs do not include such relations.

[6] We disable parameters related to content selection, syntactic template selection and lexical choice.



(a) Aggregation Operations

(b) Pragmatic Markers

Figure 2: Frequency of the Top 2 most frequent Aggregation and Pragmatic Markers in Train

simplest choice for aggregation does not combine attributes at all: this is represented by the PERIOD operator, which, if used persistently, results in an output with each content item in its own sentence as in the NO-AGG/NO-PRAG row, or the content being realized over multiple sentences as in the DISAGREEABLE row (5 sentences). However, if the other aggregation operations have a high value, PERSONAGE prefers to combine simple sentences into complex ones whenever it can, e.g., the EXTRAVERT personality example in Table 1 combines all the attributes into a single sentence by repeated use of the ALL MERGE and CONJUNCTION operations. The CONSCIENTIOUS row in Table 1 illustrates the use of the WITH-CUE aggregation operation, e.g., *with a decent rating*. Both the AGREEABLE and CONSCIENTIOUS rows in Table 1 provide examples of the ALSO-CUE aggregation operation. In PERSONAGE, the aggregation operations are defined as syntactic operations on the dictionary entry's syntactic tree. Thus to mimic these operations correctly, the neural model must derive latent representations that function as though they also operate on syntactic trees.

The pragmatic operators in the second half of Table 3 are intended to achieve particular pragmatic effects in the generated outputs: for example the use of a hedge such as *sort of* softens a claim and affects perceptions of friendliness and politeness (Brown and Levinson, 1987), while the exaggeration associated with emphasizers like *actually, basically, really* influences perceptions of extraversion and enthusiasm (Oberlander and Gill, 2004; Dewaele and Furnham, 1999). In PERSONAGE, the pragmatic parameters are attached to the syntactic tree at *insertion points* defined by syntactic constraints, e.g., EMPHASIZERS are adverbs that can occur sentence initially or before a scalar adjective. Each personality model uses a variety of pragmatic parameters. Figure 2 shows how these markers distribute differently across personality models, with examples in Table 1.

3 Model Architectures

Our neural generation models build on the open-source sequence-to-sequence (seq2seq) TGen system (Dusek and Jurcícek, 2016a)[7], implemented in Tensorflow (Abadi et al., 2016). The system is based on seq2seq generation with attention (Bahdanau et al., 2014; Sutskever et al., 2014), and uses a sequence of LSTMs (Hochreiter and Schmidhuber, 1997) for the encoder and decoder, combined with beam-search and reranking for output tuning.

The input to TGen are dialogue acts for each system action (such as *inform*) and a set of attribute slots (such as *rating*) and their values (such as *high* for attribute *rating*). The system integrates sentence planning and surface realization into a single step to produce natural language outputs. To preprocess the corpus of MR/utterance pairs, attributes that take on proper-noun values are delexicalized during training i.e., *name* and *near*. During the generation phase on the test set, a post-processing step re-lexicalizes the outputs. The MRs (and resultant embeddings) are sorted internally by dialogue act tag and attribute name.

The models are designed to systematically test the effects of increasing the level of supervision, with novel architectural additions to accommodate these changes. We use the default parameter settings from TGen (Dusek and Jurcícek, 2016a) with batch size 20 and beam size 10, and use 2,000

[7] https://github.com/UFAL-DSG/tgen

training instances for parameter tuning to set the number of training epochs and learning rate. Figure 3 summarizes the architectures.

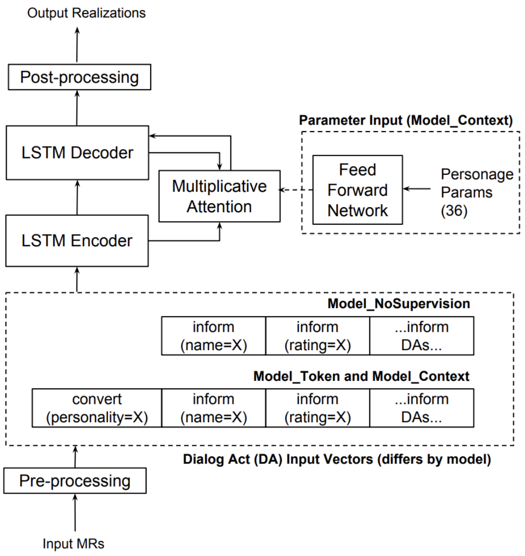


Figure 3: Neural Network Model Architecture

Model_NoSupervision. The simplest model follows the baseline TGen architecture (Dusek and Jurcícek, 2016b), with training using all 88K utterances in a single pool for up to 14 epochs based on loss monitoring for the decoder and reranker.

Model_Token. The second model adds a token of additional supervision by introducing a new dialogue act, *convert*, to encode personality, inspired by the use of a language token for machine translation (Johnson et al., 2017). Unlike other work that uses a single token to control generator output (Fan et al., 2017; Hu et al., 2017), the personality token encodes a constellation of different parameters that define the style of the matching reference. Uniquely here, the model attempts to *simultaneously* control multiple style variables that may interact in different ways. Again, we monitor loss on the validation set and training continues for up to 14 epochs for the decoder and reranker.

Model_Context. The most complex model introduces a context vector, as shown at the top right of Figure 3. The vector explicitly encodes a set of 36 style parameters from Table 3. The parameters for each reference text are encoded as a boolean vector, and a feed-forward network is added as a context encoder, taking the vector as input to the hidden state of the encoder and making the parameters available at every time step to a multiplicative attention unit. The activations of the fully connected nodes are represented as an additional time step of the encoder of the seq2seq architecture (Sutskever et al., 2014). The attention (Bahdanau et al., 2014) is computed over all of the encoder states and the hidden state of the fully connected network. Again, we set the learning rate, alpha decay, and maximum training epochs (up to 20) based on loss monitoring on the validation set.

4 Quantitative Results

Here, we present results on controlling stylistic variation while maintaining semantic fidelity.

4.1 Evaluating Semantic Quality

It is widely agreed that new evaluation metrics are needed for NLG (Langkilde-Geary, 2002; Belz and Reiter, 2006; Bangalore et al., 2000; Novikova et al., 2017a). We first present automated metrics used in NLG to measure how well model outputs compare to Personage input, then introduce novel metrics designed to fill the gap left by current evaluation metrics.

Automatic Metrics. The automatic evaluation uses the E2E generation challenge script.[8] Table 4 summarizes the results for BLEU (n-gram precision), NIST (weighted n-gram precision), METEOR (n-grams with synonym recall), and ROUGE (n-gram recall). Although the differences in metrics are small, Model_context shows a slight improvement across all of the metrics.

Model	BLEU	NIST	METEOR	ROUGE_L
NoSup	0.2774	4.2859	0.3488	0.4567
Token	0.3464	4.9285	0.3648	0.5016
Context	**0.3766**	**5.3437**	**0.3964**	**0.5255**

Table 4: Automated Metric Evaluation

Deletions, Repetitions, and Substitutions. Automated evaluation metrics are not informative about the quality of the outputs, and penalize models for introducing stylistic variation. We thus develop new scripts to automatically evaluate the types common types of neural generation errors: *deletions* (failing to realize a value), *repeats* (repeating a value), and *substitutions* (mentioning an attribute with an incorrect value).

Table 5 shows ratios for the number of deletions, repeats, and substitutions for each model for the test set of 1,390 total realizations (278 unique MRs, each realized once for each of the 5 personalities). The error counts are split by personality, and normalized by the number of unique MRs

[8] https://github.com/tuetschek/e2e-metrics

(278). Smaller ratios are preferable, indicating fewer errors. Note that because Model_NoSup does not encode a personality parameter, the error values are the same across each personality (averages across the full test set).

The table shows that Model_NoSup makes very few semantic errors (we show later that this is at the cost of limited stylistic variation). Across all error types, Model_Context makes significantly fewer errors than Model_Token, suggesting that its additional explicit parameters help avoid semantic errors. The last row quantifies whether some personalities are harder to model: it shows that across all models, DISAGREEABLE and EXTRAVERT have the most errors, while CONSCIENTIOUS has the fewest.

Model	AGREE	CONSC	DISAG	EXTRA	UNCON
Deletions					
NoSup	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
Token	0.27	0.22	0.87	0.74	0.31
Context	0.08	**0.01**	0.14	0.08	**0.01**
Repetitions					
NoSup	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Token	0.29	0.12	0.81	0.46	0.28
Context	0.02	**0.00**	0.14	**0.00**	**0.00**
Substitutions					
NoSup	0.10	0.10	0.10	0.10	0.10
Token	0.34	0.41	0.22	0.35	0.29
Context	**0.03**	**0.03**	**0.00**	**0.00**	**0.03**
All	0.68	0.35	1.96	1.29	0.61

Table 5: Ratio of Model Errors by Personality

4.2 Evaluating Stylistic Variation

Here we characterize the fidelity of stylistic variation across different model outputs.

Entropy. Shannon text entropy quantifies the amount of variation in the output produced by each model. We calculate entropy as $-\sum_{x \in S} \frac{freq}{total} * log_2(\frac{freq}{total})$, where S is the set of unique words in all outputs generated by the model, $freq$ is the frequency of a term, and $total$ counts the number of terms in all references. Table 6 shows that the training data has the highest entropy, but Model_Context performs the best at preserving the variation seen in the training data. Row NoSup shows that Model_NoSup makes the fewest semantic errors, but produces the least varied output. Model_Context, informed by the explicit stylistic context encoding, makes comparably few semantic errors, while producing stylistically varied output with high entropy.

Model	1-grams	1-2grams	1-3grams
PersonageTrain	5.97	7.95	9.34
NoSup	5.38	6.90	7.87
Token	5.67	7.35	8.47
Context	**5.70**	**7.42**	**8.58**

Table 6: Shannon Text Entropy

Pragmatic Marker Usage. To measure whether the trained models faithfully reproduce the pragmatic markers for each personality, we count each pragmatic marker in Table 3 in the output, average the counts and compute the Pearson correlation between the PERSONAGE references and the outputs for each model and personality. See Table 7 (all correlations significant with $p \leq 0.001$).

Model	AGREE	CONSC	DISAG	EXTRA	UNCON
NoSup	0.05	0.59	-0.07	-0.06	-0.11
Token	**0.35**	0.66	0.31	0.57	0.53
Context	0.28	**0.67**	**0.40**	**0.76**	**0.63**

Table 7: Correlations between Personage and models for pragmatic markers in Table 3

Table 7 shows that Model_Context has the highest correlation with the training data, for all personalities (except AGREEABLE, with significant margins, and CONSCIENTIOUS, which is the easiest personality to model, with a margin of 0.01). While Model_NoSup shows positive correlation with AGREEABLE and CONSCIENTIOUS, it shows *negative* correlation with the Personage inputs for DISAGREEABLE, EXTRAVERT, and UNCONSCIENTIOUS. The pragmatic marker distributions for Personage train in Figure 2 indicates that the CONSCIENTIOUS personality most frequently uses *acknowledgement-justify* (i.e., *"well"*, *"i see"*), and *request confirmation* (i.e., *"did you say X?"*), which are less complex to introduce into a realization since they often lie at the beginning or end of a sentence, allowing the simple Model_NoSup to learn them.[9]

Aggregation. To measure the ability of each model to aggregate, we average the counts of each aggregation operation for each model and personality and compute the Pearson correlation between the output and the PERSONAGE training data.

The correlations in Table 8 (all significant with $p \leq 0.001$) show that Model_Context has a higher correlation with Personage than the two simpler models (except for DISAGREE-

[9] We verified that there is not a high correlation between every set of pragmatic markers: different personalities do not correlate, e.g., -0.078 for Personage DISAGREEABLE and Model_Token AGREEABLE.

Model	AGREE	CONSC	DISAG	EXTRA	UNCON
NoSup	0.78	0.80	0.13	0.42	0.69
Token	0.74	0.74	**0.57**	0.56	0.60
Context	**0.83**	**0.83**	0.55	**0.66**	**0.70**

Table 8: Correlations between Personage and models for aggregation operations in Table 3

able, where Model_Token is higher by 0.02). Here, Model_NoSup actually *frequently* outperforms the more informed Model_Token. Note that *all personalities use aggregation*, even thought **not** *all personalities use pragmatic markers*, and so even without a special *personality* token, Model_NoSup is able to faithfully reproduce aggregation operations. In fact, since the correlations are frequently higher than those for Model_Token, we hypothesize that is able to more accurately focus on aggregation (common to all personalities) than stylistic differences, which Model_Token is able to produce.

5 Qualitative Analysis

Here, we present two evaluations aimed at qualitative analysis of our outputs.

Crowdsourcing Personality Judgements. Based on our quantitative results, we select Model_Context as the best-performing model and conduct an evaluation to test if humans can distinguish the personalities exhibited. We randomly select a set of 10 unique MRs from the Personage training data along with their corresponding reference texts for each personality (50 items in total), and 30 unique MRs Model_Context outputs (150 items in total).[10] We construct a HIT on Mechanical Turk, presenting a single output (either Personage or Model_Context), and ask 5 Turkers to label the output using the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003). The TIPI is a ten-item measure of the Big Five personality dimensions, consisting of two items for each of the five dimensions, one that *matches* the dimension, and one that is the *reverse* of it, and a scale that ranges from 1 (disagree strongly) to 7 (agree strongly). To qualify Turkers for the task, we ask that they first complete a TIPI on themselves, to help ensure that they understand it.

Table 9 presents results as aggregated counts for the number of times at least 3 out of the 5 Turkers rated the *matching* item for that personality higher than the *reverse* item (Ratio Correct), the average rating the correct item received (range between 1-7), and an average "naturalness" score for the output (also rated 1-7). From the table, we can see that for Personage training data, all of the personalities have a correct ratio that is higher than 0.5. The Model_Context outputs exhibit the same trend except for unconscientious and agreeable, where the correct ratio is only 0.17 and 0.50, respectively (they also have the lowest correct ratio for the original Personage data).

	Personage			Model_Context		
Person.	Ratio Correct	Avg. Rating	Nat. Rating	Ratio Correct	Avg. Rating	Nat. Rating
agree	0.60	4.04	5.22	0.50	4.04	4.69
disagr	0.80	4.76	4.24	0.63	4.03	4.39
consc	1.00	5.08	5.60	0.97	5.19	5.18
uncon	0.70	4.34	4.36	0.17	3.31	4.58
extra	0.90	5.34	5.22	0.80	4.76	4.61

Table 9: Percentage of Correct Items and Average Ratings and Naturalness Scores for Each Personality (Personage vs. Model_Context)

Table 9 also presents results for naturalness for both the reference and generated utterances, showing that both achieve decent scores for naturalness (on a scale of 1-7). While human utterances would probably be judged more natural, it is not at all clear that similar experiments could be done with human generated utterances, where it is difficult to enforce the same amount of experimental control.

Generalizing to Multiple Personalities. A final experiment explores whether the models learn additional stylistic generalizations not seen in training. We train a version of Model_Token, as before on instances with single personalities, but such that it can be used to generate output with a combination of *two* personalities. The experiment uses the original training data for Model_Token, but uses an expanded test set where the MR includes **two** personality convert tags. We pair each personality with all personalities except its exact opposite.

Sample outputs are given in Table 10 for the disagreeable personality, which is one of the most distinct in terms of aggregation and pragmatic marker insertion, along with occurrence counts (frequency shown scaled down by 100) of the operations that it does most frequently: specifically, *period aggregation* and *expletive pragmatic markers*. Rows 1-2 shows the counts and an exam-

[10] Note that we use fewer Personage references simply to validate that our personalities are distinguishable in training, but will more rigorously evaluate our model in future work.

	Persona	Period Agg.	Explet Prag.	Example
1	DISAG	5.71	2.26	Browns Cambridge is damn moderately priced, also it's in city centre. It is a pub. It is an italian place. It is near Adriatic. It is damn family friendly.
2	CONSC	0.60	0.02	Let's see what we can find on Browns Cambridge. I see, well it is a pub, also it is moderately priced, an italian restaurant near Adriatic and family friendly in city centre.
3	DISAG+ CONSC	3.81	0.84	Browns Cambridge is an italian place and moderately priced. It is near Adriatic. It is kid friendly. It is a pub. It is in riverside.

Table 10: Multiple-Personality Generation Output based on DISAGREEABLE

ple of each personality on its own. The combined personality output is shown in Row 3. We can see from the table that while CONSCIENTIOUS on its own realizes the content in two sentences, period aggregation is much more prevalent in the DISAGREEABLE + CONSCIENTIOUS example, with the same content being realized in 5 sentences. Also, we see that some of the expletives originally in DISAGREEABLE are dropped in the combined output. This suggests that the model learns a combined representation unlike what it has seen in train, which we will explore in future work.

6 Related Work and Conclusion

The restaurant domain has long been a testbed for conversational agents with much earlier work on NLG (Howcroft et al., 2013; Stent et al., 2004; Devillers et al., 2004; Gašic et al., 2008; Mairesse et al., 2010; Higashinaka et al., 2007), so it is not surprising that recent work using neural generation methods has also focused on the restaurant domain (Wen et al., 2015; Mei et al., 2015; Dusek and Jurcícek, 2016b; Lampouras and Vlachos, 2016; Juraska et al., 2018). The restaurant domain is ideal for testing generation models because sentences can range from extremely simple to more complex forms that exhibit discourse relations such as justification or contrast (Stent et al., 2004). Most recent work focuses on achieving semantic fidelity for simpler syntactic structures, although there has also been a focus on crowdsourcing or harvesting training data that exhibits more stylistic variation (Novikova et al., 2017; Nayak et al., 2017; Oraby et al., 2017).

Most previous work on neural stylistic generation has been carried out in the framework of "style transfer": this work is hampered by the lack of parallel corpora, the difficulty of evaluating content preservation (semantic fidelity), and the challenges with measuring whether the outputs realize a particular style. Previous experiments attempt to control the sentiment and verb tense of generated movie review sentences (Hu et al., 2017), the content preservation and style transfer of news headlines and product review sentences (Fu et al., 2018), multiple automatically extracted style attributes along with sentiment and sentence theme for movie reviews (Ficler and Goldberg, 2017), sentiment, fluency and semantic equivalence (Shen et al., 2017), utterance length and topic (Fan et al., 2017), and the personality of customer care utterances in dialogue (Herzig et al., 2017). However, to our knowledge, no previous work evaluates simultaneous achievement of multiple targets as we do. Recent work introduces a large parallel corpus that varies on the formality dimension, and introduces several novel evaluation metrics, including a custom trained model for measuring semantic fidelity (Rao and Tetreault).

Other work has also used context representations, but not in the way that we do here. In general, these have been used to incorporate a representation of the prior dialogue into response generation. Sordoni et al. (2015) propose a basic approach where they incorporate previous utterances as a bag of words model and use a feed-forward neural network to inject a fixed sized context vector into the LSTM cell of the encoder. Ghosh et al. (2016) proposed a modified LSTM cell with an additional gate that incorporates the previous context as input during encoding. Our context representation encodes stylistic parameters.

This paper evaluates the ability of different neural architectures to faithfully render the semantic content of an utterance while simultaneously exhibiting stylistic variations characteristic of Big Five personalities. We created a novel parallel training corpus of over 88,000 meaning representations in the restaurant domain, and matched reference outputs by using an existing statistical natural language generator, PERSONAGE (Mairesse and Walker, 2010). We design three neural models that systematically increase the stylistic encodings given to the network, and show that MODEL_CONTEXT benefits from the greatest explicit stylistic supervision, producing outputs that both preserve semantic fidelity and exhibit distinguishable personality styles.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*.

Elisabeth André, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes. 2000. The automated design of believable dialogues for animated presentation teams. *Embodied conversational agents* pages 220–255.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In ICLR*.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proc. of the First International Natural Language Generation Conference (INLG2000)*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *EACL*.

Douglas Biber. 1991. Variation across speech and writing Cambridge University Press.

Penelope Brown and Steve Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Ondrej Dusek and Filip Jurcícek. 2016a. A context-aware natural language generator for dialogue systems. *In SIGDIAL* pages 85-190.

Ondrej Dusek and Filip Jurcícek. 2016b. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* pages 45-51.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. *EACL 2014* page 702.

Laurence Devillers, Hélène Maynard, Sophie Rosset, Patrick Paroubek, Kevin McTait, Djamel Mostefa, Khalid Choukri, Laurent Charnay, Caroline Bousquet, Nadine Vigouroux, et al. 2004. The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

Jean-Marc Dewaele and Adrian Furnham. 1999. Extraversion: the unloved variable in applied linguistic research. *Language Learning*, 49(3):509–544.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *CoRR* abs/1711.05217.

Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proc. of the Workshop on Stylistic Variation at EMNLP 18*. pages 94–104.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 663–670.

M. Gašic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, K. Yu, and S. Young. 2008. Training and evaluation of the his-pomdp dialogue system in noise. *Proc. Ninth SIGdial, Columbus, OH* .

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291* .

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. 2003. A very brief measure of the big five personality domains. *Journal of Research in Personality* Vol. 37:504–528.

Jonathan Herzig and Michal Shmueli-Scheuer and Tommy Sandbank and David Konopnicki. 2017. Neural Response Generation for Customer Service based on Personality Traits. In *Proc. of the INLG*.

Ryuichiro Higashinaka, Marilyn A. Walker, and Rashmi Prasad. 2007. An unsupervised method for learning generation dictionaries for spoken dialogue systems by mining user reviews. *ACM Transactions on Speech and Language Processing* 4(4).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

David M Howcroft, Crystal Nakatsu, and Michael White. 2013. Enhancing the expression of contrast in the sparky restaurant corpus. *ENLG 2013* page 30.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Towards controlled generation of text. *International Conference on Machine Learning* pages 1587–1596.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: enabling zero-shot translation. *In Transactions of the Association for Computational Linguistics* pages 339-351.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proc. of*

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pages 638–646.

Juraj Juraska and Panagiotis Karagiannis and Kevin Bowden and Marilyn Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proc. of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Gerasimos Lampouras and Andreas Vlachos. 2016. Imitation learning for language generation from unaligned data. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING*. ACL, pages 1101–1112.

I. Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence gene rator. In *Proc. of the INLG*.

Benoit Lavoie and Owen Rambow. 1997. A fast and portable realizer for text generation systems. In *Proc. of the Third Conference on Applied Natural Language Processing, ANLP97*. pages 265–268.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* pages 994-1003.

F. Mairesse and M.A. Walker. 2011. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics Journal, Vol. 37 Issue 3* pages 455–488.

F. Mairesse and M.A. Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction* pages 1–52.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* pages 1552–1561.

François Mairesse and Marilyn A. Walker. 2008. Trainable generation of Big-Five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proc. of Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 720–730.

Igor A. Melčuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY, Albany, New York.

Neha Nayak, Dilek Hakkani-Tur, Marilyn Walker, and Larry Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *Proc. of Interspeech 2017*.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue* pages 201-206.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *International Conference on Natural Language Generation*.

J. Oberlander and A. Gill. 2004. Individual differences and implicit language: personality, parts-of-speech, and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040.

Shereen Oraby, Sheideh Homayon, and Marilyn Walker. 2017. Harvesting Creative Templates for Generating Stylistically Varied Restaurant Reviews. In *Proc. of the Workshop on Stylistic Variation at EMNLP 18*. pages 28–36.

James W. Pennebaker, L. E. Francis, and R. J. Booth. 2001. *LIWC: Linguistic Inquiry and Word Count*.

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue. 1992. Experiments in evaluating interactive spoken language systems. In *Proc. of the DARPA Speech and NL Workshop*. pages 28–33.

Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proc. of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in Neural Information Processing Systems* pages 6833–6844.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015.

A neural network approach to context-sensitive generation of conversational responses. In *Proc. of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 196–205.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialogue systems. In *Meeting of the Association for Computational Linguistics*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Marilyn Walker, Rashmi Prasad, and Amanda Stent. 2003. A trainable generator for recommendations in multimodal dialog. In *EUROSPEECH*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* pages 1711–1721.

Steve Whittaker, Marilyn Walker, and Johanna Moore. 2002. Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. In *Language Resources and Evaluation Conference*.